Methods for bandwidth detection in kernel conditional density estimations

Kateřina Konečná*[1] and Ivana Horová[2]

[1,2]*Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic.*

This contribution is focused on the kernel conditional density estimations (KCDE). The estimation depends on the smoothing parameters which influence the final density estimation significantly. This is the reason why a requirement of any data-driven method is needed for bandwidth estimation. In this contribution, the cross-validation method, the iterative method and the maximum likelihood approach are conducted for bandwidth selection of the estimator. An application on a real data set is included and the proposed methods are compared.

Keywords: kernel conditional density estimation, bandwidth detection, cross-validation method, iterative method, maximum likelihood method

Introduction

Kernel smoothing techniques belong to the most popular non-parametric techniques for data interpolation, especially for its simple usage and no strictly limiting requirements. Conditional density estimations offer the comprehensive information about the data structure – regression models only the conditional expectation while conditional density includes even the variability and the whole data distribution.
The estimator depends on the unknown parameters, called the smoothing parameters or bandwidths. They influence the quality of the estimation significantly, this is the reason why so much attention is given to the bandwidth determination. The optimal values of the smoothing parameters depend on the unknown conditional and marginal density, thus there is a necessity to develop an automatic data-driven bandwidth selectors. In this contribution, the widely used cross-validation method is supplemented with the iterative method and the leave-one-out maximum likelihood method.

*Corresponding author: xkonecn3@math.muni.cz

1 Statistical properties of the Nadaraya-Watson estimator of conditional density

The basic building block of kernel smoothing is a kernel function, which plays a role of weighting function. Let K be a real valued function satisfying

1. $K \in \mathrm{Lip}[-1,1]$, i. e. $|K(x) - K(y)| \le L|x-y|$, $\forall x, y \in [-1,1]$, $L > 0$,
2. $\mathrm{supp}(K) = [-1,1]$,
3. moment conditions:
$$\int_{-1}^{1} K(x)\,\mathrm{d}x = 1,\ \int_{-1}^{1} xK(x)\,\mathrm{d}x = 0,\ \int_{-1}^{1} x^2K(x)\,\mathrm{d}x = \beta_2(K) \neq 0.$$

Such a function K is called a kernel of order 2.
Conditional density models the probability of a random variable Y given a fixed observation $X = x$. The Nadaraya-Watson estimator of conditional density takes the form

$$\hat{f}_{NW}(y|x) = \frac{1}{h_y}\sum_{i=1}^{n} w_i^{NW}(x) K\left(\frac{y-Y_i}{h_y}\right), \tag{1}$$

where $w_i^{NW}(x) = \dfrac{K\left(\frac{x-X_i}{h_x}\right)}{\sum\limits_{j=1}^{n} K\left(\frac{x-X_j}{h_x}\right)}$ is a weight function in the point x, $h_x, h_y > 0$ are the smoothing parameters.

The statistical properties of the estimator are the rudiments for appraisal of suitability of the estimator and determination of the optimal values of bandwidths.
The Asymptotic Bias (AB) and the Asymptotic Variance (AV) of the Nadaraya-Watson estimator are given by Hyndman *et al.* ([4]) with the expressions

$$\mathrm{AB}\left\{\hat{f}_{NW}(y|x)\right\} = \frac{1}{2}h_x^2\beta_2(K)\left[2\frac{g'(x)}{g(x)} + \frac{\partial^2 f(y|x)}{\partial x^2}\right] + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2},$$
$$\mathrm{AV}\left\{\hat{f}_{NW}(y|x)\right\} = \frac{R^2(K)f(y|x)}{nh_xh_yg(x)},$$

where $R(K) = \int K^2(t)\,\mathrm{d}t$, $g(x)$ is a marginal density of a random variable X. The global quality of the estimate is measured by the Mean Integrated Squared Error (MISE) in the form

$$\mathrm{MISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\} = \iint \mathrm{E}\left\{\left(\hat{f}_{NW}(y|x) - f(y|x)\right)^2\right\} g(x)\,\mathrm{d}x\,\mathrm{d}y.$$

The main term of MISE $\left\{\hat{f}_{NW}(\cdot|\cdot)\right\}$, the Asymptotic Mean Integrated Squared Error (AMISE), is of the form

$$\text{AMISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\} = \frac{c_1}{nh_xh_y} + c_2h_x^4 + c_3h_y^4 + c_4h_x^2h_y^2,$$

where

$$\begin{aligned}
c_1 &= \int R^2(K)\,\mathrm{d}x,\\
c_2 &= \iint \frac{\beta_2^2(K)}{4}\left(2\frac{g'(x)}{g(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 g(x)\,\mathrm{d}y\,\mathrm{d}x,\\
c_3 &= \iint \frac{\beta_2^2(K)}{4}\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2 g(x)\,\mathrm{d}y\,\mathrm{d}x,\\
c_4 &= \iint \frac{\beta_2^2(K)}{2}\left(2\frac{g'(x)}{g(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2}\right)\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)g(x)\,\mathrm{d}y\,\mathrm{d}x.
\end{aligned}$$

The optimal bandwidths $\left(h_x^*, h_y^*\right)$ minimize AMISE

$$(h_x^*, h_y^*) = \underset{(h_x,h_y)}{\arg\min}\,\text{AMISE}\left\{\hat{f}_{NW}(\cdot|\cdot)\right\},$$

where the nonequations $an^{-1/6} \leq h_x \leq bn^{-1/6}$ and $cn^{-1/6} \leq h_y \leq dn^{-1/6}$ are held for $0 < a < b < \infty$ and $0 < c < d < \infty$. The optimal values of smoothing parameters are derived by differentiating of AMISE, setting the derivatives to 0 and making several algebraic simplifications. They are given by Hyndman *et al.* in the paper [4] as follows

$$\begin{aligned}
h_x^* &= n^{-1/6}c_1^{1/6}\left[4\left(\frac{c_3^5}{c_4}\right)^{1/4} + 2c_5\left(\frac{c_3}{c_4}\right)^{3/4}\right]^{-1/6},\\
h_y^* &= h_x^*\left(\frac{c_3}{c_4}\right)^{1/4} = n^{-1/6}c_1^{1/6}\left[4\left(\frac{c_4^5}{c_3}\right)^{1/4} + 2c_5\left(\frac{c_4}{c_3}\right)^{3/4}\right]^{-1/6}.
\end{aligned}$$

2 Methods for bandwidth detection

The optimal values of the smoothing parameters depend on the unknown conditional and marginal density. This is the reason why any data-driven method for the estimation of them is needed.

One of the most common methods for choosing the bandwidths is the **cross-validation method** introduced by Fan and Yim [2] and Hansen [3]. The idea of the method consists in minimization of the proper estimate of the Integrate Squared Error (ISE) represented by the cross-validation function

$$CV\left(h_x, h_y\right) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i,NW}\left(y|X_i\right)^2 \,\mathrm{d}y - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,NW}\left(Y_i|X_i\right),$$

where $\hat{f}_{-i,NW}\left(y|x\right)$ is the estimate in the pair of points $\left(X_i, Y_i\right)$ using the points $\left\{\left(X_j, Y_j\right), j \neq i\right\}$. Thus, the estimates of bandwidths are given by

$$(\hat{h}_x^{CV}, \hat{h}_y^{CV}) = \underset{(h_x,h_y)}{\arg\min}\, CV(h_x, h_y).$$

The next proposed method is the **iterative method** suggested by Konečná and Horová ([5]). The method is based on a suitable estimation of AMISE which can be expressed by a sum of the Asymptotic Integrated Variance (AIV) and the Asymptotic Integrated Squared Bias (AISB). The relation (2) is derived by differentiating of AMISE, setting the derivatives to 0, and by replacing the terms by their estimations:

$$\mathrm{AIV}\left\{\hat{f}(\cdot|\cdot)\;\right\} - 2\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\;\right\} = 0. \tag{2}$$

The term $\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\right\}$ is an approximation of the $\mathrm{AISB}\left\{\hat{f}(\cdot|\cdot)\right\}$ term and it is of the form

$$\begin{aligned}\widehat{\mathrm{ISB}}\left\{\hat{f}(\cdot|\cdot)\right\} &= \iint \left(\widehat{\mathrm{bias}}\left\{\hat{f}(y|x)\right\}\right)^2 g(x)\,\mathrm{d}x\,\mathrm{d}y\\ &= \iint \left(\frac{\sum\limits_i K_{h_x\sqrt{2}}\left(x - X_i\right) K_{h_y\sqrt{2}}\left(y - Y_i\right)}{\sum\limits_i K_{h_x\sqrt{2}}\left(x - X_i\right)} - \hat{f}_{NW}(y|x)\right)^2 g(x)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

The supplemented equation $\hat{h}_y = \hat{c}\hat{h}_x$ to the equation (2) is represented by a relation $\hat{c}$ between the values of the smoothing parameters, $\hat{c}$ is given by the reference rule suggested by Bashtannyk and Hyndman in the paper [1]. The estimations of the smoothing parameters are derived as a solution of the system of two nonlinear equations (2) and the equation $\hat{h}_y = \hat{c}\hat{h}_x$.

The last suggested method is the **leave-one-out maximum likelihood method** which proceeds with the maximum likelihood method, a statistical standard procedure for estimating the unknown parameters. We consider a random vector $(\mathbf{X}, \mathbf{Y})$ with the independent and identically distributed observations $(X_i, Y_i), i = 1, \ldots, n$ of the unknown conditional density. We define the modified likelihood function

$$\mathcal{L}\left(h_x, h_y \mid \mathbf{X}, \mathbf{Y}\right) = \prod_{j=1}^{n} \hat{f}_{-j,NW}\left(Y_j \mid X_j\right).$$

The modification of the classical likelihood approach consists in leaving one observation out. The estimations of the smoothing parameters are given by maximization of $\mathcal{L}\left(h_x, h_y \mid \mathbf{X}, \mathbf{Y}\right)$, i.e.

$$(\hat{h}_x^{\mathcal{L}}, \hat{h}_y^{\mathcal{L}}) = \arg\min_{(h_x, h_y)} \mathcal{L}\left(h_x, h_y \mid \mathbf{X}, \mathbf{Y}\right).$$

3 Application on a real data

For comparison of the proposed methods, the `airquality` data from the `datasets` package in R ([6]) are concerned. The data describe daily air quality in New York, May to September 1973. The estimation of mean ozone concentration in parts per billion, given the maximum daily temperature in degrees Fahrenheit is focused on. There is 153 observations in total, in fact, we include only 116 observation because of some missing values.
The cross-validation method (CV), the iterative method (IT) and the leave-one-out maximum likelihood (ML) are used for bandwidth detection. The values of estimated bandwidths and the computational times are given in the Table 1.

Table 1: Estimates of the smoothing parameters and computational times for methods used for bandwidth determination.

method	$\hat{h}_x$	$\hat{h}_y$	computational time $[s]$
CV	1.845	7.638	182
IT	6.289	15.276	67.6
ML	2.517	10.017	31.3

As it can be seen, the CV method gives the most undersmoothed estimation due to small values of the smoothing parameters, whereas the IT method gives the most oversmoothed estimation. It seems that the ML gives the best results, supported by the shortest computational time. The IT method is the fastest – it takes about 30 seconds, the computational difficulty of the other two methods is evident. The IT method takes less than one third while the

ML method takes even one sixth of the CV's computational time.

It is important to emphasize that these results are valid for this real-data application. Several simulation studies should be executed for the proper assessment of the proposed methods, although this exceeds the extent of this contribution.

4 Conclusion

In this contribution, the methods for bandwidth determination were focused on. The classical approach for bandwidth detection, the cross-validation method, was supplemented with two suggested methods – the iterative and the leave-one-out maximum likelihood method.

These approaches could be extended to the other types of kernel conditional density estimations which have not been mentioned in this contribution. Future work could also involve variable bandwidths, on the other hand, their theoretical aspect as well as computational implementation would be quite difficult.

Acknowledgements: The research was supported by the Czech Science Foundation no. GA15-06991S.

References

[1] D. M. Bashtannyk and R. J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279 – 298, 2001.

[2] J. Fan and T. H. Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

[3] B. E. Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004.

[4] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996.

[5] K. Konečná and I. Horová. *Conditional Density Estimations*, pages 15–31. International Society for the Advancement of Science and Technology (ISAST), Athens, 2014.

[6] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.